Att's Docket No.: 11646-011001

APPLICATION

FOR

UNITED STATES LETTERS PATENT

TITLE: DIRECT NAVIGATION FOR INFORMATION RETRIEVAL

APPLICANT: JANE WEN CHANG AND KENNEY NG

CERTIFICATE OF MAILING BY EXPRESS MAIL

Express Mail Label No. EL486014504US

I hereby certify under 37 CFR §1.10 that this correspondence is being deposited with the United States Postal Service as Express Mail Post Office to Addressee with sufficient postage on the date indicated below and is addressed to the Commissioner for Patents, Washington, D.C. 20231.

February 22, 2002
Date of Deposit

Signature

Lervia Jenkins
Typed or Printed Name of Person Signing Certificate

Direct Navigation for Information Retrieval

BACKGROUND

This invention relates to direct navigation for information retrieval.

A web page such as on the World Wide Web ("Web") represents web content and is typically written in Hypertext Markup Language ("HTML"). HTML is a set of markup symbols or codes inserted in a file intended for display on a Web browser page. The markup symbols tell the Web browser how to display a Web page's words and images for a user.

A search engine is Web-enabled software that receives search terms, i.e., a query, from a user and identifies documents, i.e., web pages, on the Web that is otherwise associated with, the search terms. The documents are typically identified by uniform resource locators ("URLs") that are links to their content.

SUMMARY

In an aspect, the invention features a method of retrieving information including assigning concept labels to documents contained in a collection, receiving a query, converting the query to a query concept, and mapping the query concept to a concept label.

Embodiments may include one or more of the following. Assigning may include parsing the documents automatically with a grammar. The concept label may represent a general notion. The query may be a text query received from a user. Assigning may include spidering the collection, matching features contained in each of the documents to a store of concepts, and storing document location indicators for each matched concept. The documents may be HyperText Markup Language (HTML) files. The document location indicators may be Universal Resource Identifiers (URLs). Converting may include applying a store of grammar rules to the query and the grammar rules may map text to concepts.

The method may also include generating a list of the mapping in which the list represents locations of documents. The locations may be Universal Resource Identifiers (URLs).

In another aspect, the invention features a method of document retrieval including assigning concept labels to documents contained in a collection according to grammar rules, receiving a query, converting the query to a query concept using the grammar rules, and mapping the query concept to a concept label.

Embodiments may include one or more of the following. Assigning may include parsing the documents automatically with the grammar rules. The query may be received from a user.

The method may also include generating a list of the mapped query concepts, and displaying the list to the user on an input/output device.

Embodiments of the invention may have one or more of the following advantages.

Documents in a collection are pre-labeled with concepts, wherein each concept is a general notion or idea. A grammar is written around concepts. The grammar is applied to a user query and provides a direct mapping of the user query to appropriate documents in a collection of documents found on the Web.

The pre-labeling with concepts may be automatic using parsing with a grammar.

A user query is responded to using concept matching rather than direct word matching.

Using a grammar applied to a user query provides a technique for allowing many different ways of expressing something to map to a single item.

Annotating documents contained on the Web with concepts provides a robust manner of searching for Web documents.

Concept matching overcomes the limitation where a user query needs to match words and cannot find all the words in a single document.

DESCRIPTION OF DRAWINGS

The foregoing features and other aspects of the invention will be described further in detail by the accompanying drawings, in which:

FIG. 1 is a block diagram of a network.

FIG. 2 is a block diagram of a direct navigation process.

FIG. 3 is a flow diagram of a query process.

DETAILED DESCRIPTION

Referring to FIG. 1, an exemplary network 10 includes a user system 12 linked to a globally connected network of computers such as the Internet 14. The network 10 includes content servers 16, 18 and 20 linked to the Internet 14. Although only one user system 12 and three content server systems 16, 18, 20 are shown, other configurations include numerous client systems and numerous server systems. Each content server system 16, 18, 20 includes a corresponding storage device 22, 24, 26. Each storage device 22, 24, 26 includes a corresponding database of content 28, 30, 32. The network 10 also includes a direct navigation server 34.

The direct navigation server 34 includes a processor 36 and a memory 38. Memory 38 stores an operating system ("O/S") 40, a TCP/IP stack 42 for communicating over the network 10, and machine-executable instructions 44 executed by processor 36 to

perform a direct navigation process 100 described below. The direct navigation server 34 also includes a storage device 46 having a database 47.

The user system 12 includes an input/output device 48 having a Graphical User Interface (GUI) 50 for display to a user 52. The GUI 50 typically executes search engine software such as the Yahoo, AltaVista, Lycos or Goggle search engine through browser software, such as Netscape Communicator from AOL Corporation or Internet Explorer from Microsoft Corporation.

Referring to FIG. 2, the direct navigation process 100 includes a pre-processing stage 102 and a post-processing stage 104. The pre-processing stage 102 includes a document annotation process 106. The post-processing stage 104 includes a user query conversion process 108 and a concept mapping process 110.

The annotation process 106 assigns one or more concept labels on features of a document contained in a collection of documents. A concept label represents a concept and a concept represents a general notion or idea. More specifically, each of the databases 28, 30, 32 contain pages (e.g., web pages) generally referred to as documents. The annotation process 106 includes a spider program to spider all the pages of content contained in the databases 28, 30, 32.

A spider is a program that visits Web sites (e.g., server 16, 18, 20) and reads their pages and other information in order to generate entries for a search engine index. The major search engines on the Web all have such a program, which is also known as a "crawler" or a "bot." Spiders are typically programmed to visit servers (i.e., websites) that have been submitted by their owners as new or updated. Entire Web sites or specific pages can be selectively visited and indexed. Spiders are called spiders because they usually visit many Web sites in parallel at the same time, their "legs" spanning a large area of the "web." Spiders can crawl through a site's pages in several ways. One way is to follow all the hypertext links in each page until all the pages have been read. Hypertext is an organization of information units into connected associations that a user can choose to make. An instance of such an association is called a link or hypertext link. A link is a selectable connection from one word, picture, or information object to another. In a multimedia environment such as the World Wide Web, such objects can include sound and motion video sequences. The most common form of link is the highlighted word or picture that can be selected by the user (with a mouse or in some other fashion), resulting in the immediate delivery and view of another file. The highlighted object is referred to as an anchor. The anchor

reference and the object referred to constitute a hypertext link.

For example, a particular page of content may include text pertaining to automobiles, with their associated purchase options and pricing. This particular page can be annotated with a "review" concept and/or an "automotive review" concept. The document annotation process 106 stores a Universal Resource Identifier (URL) of the particular page along with its related concept(s) in the storage device 46 of the direct navigation server 34. The URL is an address of the particular page (also referred to as a resource). The type of resource depends on the Internet application protocol. Using the World Wide Web's protocol, the Hypertext Transfer Protocol (HTTP), the resource can be an HTML page, an image file, a program such as a common gateway interface application or Java applet, or any other file supported by HTTP. The URL has the name of the protocol required to access the resource, a domain name that identifies a specific computer on the Internet, and a hierarchical description of a page location on the computer.

Concepts can be generated manually and matched to the particular web page. Concepts can also be generated automatically by parsing the documents with a grammar. In another example, concepts can be generated from a review of features associated with the page being spidered.

In the post-processing stage 104, the user query conversion process 108 receives text in a user query entered by the user 52 through search engine software executing through browser software. In another example, the process 108 receives audio input as the user query and converts the audio input into a text user query.

The user query conversion process 108 utilizes a set of grammar rules stored in the storage device 46 of the direct navigation server 32 and applies the grammar rules to the user query such that the query matches one or more concepts.

The user query may be a word or multiple words, sentence fragments, a complete sentence, and may contain punctuation. The query is normalized as pretext. Normalization includes checking the text for spelling and proper separation. A language lexicon is also consulted during normalization. The language lexicon specifies a large list of words along with their normalized forms. The normalized forms typically include word stems only, that is, the suffixes are removed from the words. For example, the word "computers" would have the normalized form "computer" with the plural suffix removed.

The normalized text is parsed, converting the normalized text into fragments adapted for further processing. Annotating words as punitive keys and values, according to a feature lexicon, produces fragments. The feature lexicon is a

vocabulary, or book containing an alphabetical arrangement of the words in a language or of a considerable number of them, with the definition of each e.g., a dictionary. For example, the feature lexicon may specify that the term "Compaq" is a potential value and that "CPU speed" is a potential key. Multiple annotations are possible.

The fragments are inflated by the context in which the text inputted by the user arrived, e.g., a previous query, if any, that was inputted and/or a content of a web page in which the user text was entered. The inflation is preformed by selectively merging state information provided by a session service with a meaning representation for the current query. The selective merging is configurable based on rules that specify which pieces of state information from the session service should be merged into the current meaning representation and which pieces should be overridden or masked by the current meaning representation.

A session service may store all of the "conversations" that occur at any given moment during all of the user's session. State information is stored in the session service providing a method of balancing load with additional computer configurations. Load balancing may send each user query to a different configuration of the computer system. However, since query processing requires state information, storage of station

information on the computer system will not be compatible with load balancing. Hence, use of the session service provides easy expansion by the addition of server systems, with load sharing among the systems to support more users.

The state information includes user-specified constraints that were used in a previous query, if any. The state information may optionally include a result set, either in its entirety or in condensed form, from the previous query to speed up subsequent processing in context. The session service may reside in one computer system, or include multiple computer systems. When multiple computer systems are employed, the state information may be assigned to a single computer system or replicated across more than one computer system.

The inflated sentence fragments are converted into meaning representation by making multiple passes through a meaning resolution stage. The meaning resolution stage determines if there is a valid interpretation within the text query of a key-value grouping of the fragment. If there is a valid interpretation, the key value grouping is used. For example, if the input text, i.e., inflated sentence fragment, contains the string "500 MHz CPU speed," which may be parsed into two fragments, "500 MHz" value and "CPU speed" key, then there is a valid grouping of key = "CPU speed" and value = "500 MHz".

If no valid interpretation exists, a determination is made on whether the grammar rules contain a valid interpretation. If there is a valid interpretation, the key value group is used. If no valid interpretation is found, a determination of whether previous index fields have a high confidence of uniquely containing the fragment. If so, the key value grouping is used. If not, other information sources are searched and a valid key value group generated. If a high confidence and valid punitive key is determined through one of the information sources consulted, then the grouping of the key and value form an atomic element are used. To make it possible to override false interpretations, a configuration of grammar can also specify manual groupings of keys and values that take precedence over the meaning resolution stage. Meaning resolved fragments, representing the user query, are associated with concepts.

The concept mapping process 110 matches the concept associated resolved fragments to concept/URL pairs stored in the storage device 46 and loads the associated URL representing the matched concepts.

Referring to FIG. 3, a query process 200 includes loading (202) a database containing pages of content with mapped concepts. The process 200 receives (204) a user query and parses (206) the user query in conjunction with grammar rules. The process 200 associates (208) the parsed user query with a

concept. The process 200 matches (210) the user query concept with a concept/URL pair and loads (212) the associated concept as directed by the URL.

Accordingly, other embodiments are within the scope of the following claims.